# Scaling Laws for Autoregressive Generative Modeling

Year: 2020 | Citations: 537 | Authors: T. Henighan, J. Kaplan, Mor Katz, Mark Chen, Christopher Hesse

---

## Abstract

We identify empirical scaling laws for the cross-entropy loss in four domains: generative image modeling, video modeling, multimodal image$\leftrightarrow$text models, and mathematical problem solving. In all cases autoregressive Transformers smoothly improve in performance as model size and compute budgets increase, following a power-law plus constant scaling law. The optimal model size also depends on the compute budget through a power-law, with exponents that are nearly universal across all data domains.
The cross-entropy loss has an information theoretic interpretation as $S($True$) + D_{\mathrm{KL}}($True$||$Model$)$, and the empirical scaling laws suggest a prediction for both the true data distribution's entropy and the KL divergence between the true and model distributions. With this interpretation, billion-parameter Transformers are nearly perfect models of the YFCC100M image distribution downsampled to an $8\times 8$ resolution, and we can forecast the model size needed to achieve any given reducible loss (ie $D_{\mathrm{KL}}$) in nats/image for other resolutions.
We find a number of additional scaling laws in specific domains: (a) we identify a scaling relation for the mutual information between captions and images in multimodal models, and show how to answer the question "Is a picture worth a thousand words?"; (b) in the case of mathematical problem solving, we identify scaling laws for model performance when extrapolating beyond the training distribution; (c) we finetune generative image models for ImageNet classification and find smooth scaling of the classification loss and error rate, even as the generative loss levels off. Taken together, these results strengthen the case that scaling laws have important implications for neural network performance, including on downstream tasks.